ON RESTRICTED SUBSET SELECTION RULES FOR SELECTING THE BEST POPULATION*

by

Shanti S. Gupta
Department of Statistics
Purdue University
West Lafayette, IN 47907-1399

and

TaChen Liang
Department of Mathematics
Wayne State University
Detroit, MI 48202

Technical Report # 91-27C

Department of Statistics
Purdue University

May, 1991

*This research was supported in part by NSF Grants DMS–8923071 and DMS–8717799 at Purdue University.

△

ON RESTRICTED SUBSET SELECTION RULES FOR SELECTING THE BEST POPULATION

by

Shanti S. Gupta
Department of Statistics
Purdue University
West Lafayette, IN 47907-1399

and

TaChen Liang
Department of Mathematics
Wayne State University
Detroit, MI 48202

Abstract

This paper deals with the problem of selecting the best from among $K(\geq 2)$ location parameter models having cdf $G(x-\theta_i), i=1,\ldots,k$, respectively. A population π_i is said to be the best if $\theta_i = \max_{1\leq j\leq k} \theta_j$. For a specified positive constant δ^*, a population π_i is said to be good if $\max_{1\leq j\leq k} \theta_j - \theta_i \leq \delta^*$, and bad otherwise. We assume that there is no prior information about the possible configurations of the parameters $\theta_1,\ldots,\theta_k$. Our goal is to select a subset so that the best population is included in the selected subset, and only good populations are selected. A selection procedure achieving the P^*-condition for the general location-parameter models is proposed. We then specialize it for normal distribution models $N(\theta_i,\sigma^2)$. Some modified selection rules are also investigated. These modified selection rules will achieve the P^*-condition as well as control the expected value of the number of bad populations selected. Finally, an example is presented to illustrate the implementation of the selection rules.

AMS 1980 Subjective Classification: 62F07

Keywords and phrases: Best population; P^*-condition; selection goal; subset selection; correct selection; two-stage selection rule; location-parameter.

1. Introduction

Consider $k(\geq 2)$ independent location-parameter models $\pi_1, \ldots, \pi_k$, which have absolutely continuous cumulative distribution functions (cdf) $G(x-\theta_1), \ldots, G(x-\theta_k)$, respectively, where $-\infty < \theta_i < \infty, i = 1, \ldots, k$. Let $\underset{\sim}{\theta} = (\theta_1, \ldots, \theta_k)$ and let $\theta_{[1]} \leq \ldots \leq \theta_{[k]}$ denote the ordered values of $\theta_1, \ldots, \theta_k$. It is assumed that the exact pairing between the ordered parameters and the unordered parameters is unknown. The population associated with the largest location parameter $\theta_{[k]}$ is called the best population. In many practical situations, the experimenter is interested in the selection of the best population.

The problem of selecting the best population from among k populations has been studied extensively in the literature. A lot of selection procedures have been derived for different selection goals by several authors. When the underlying populations have normal distributions $N(\theta_i, \sigma^2)$, Bechhofer (1954) introduced the indifference zone approach by assuming the restriction $\theta_{[k]} - \theta_{[k-1]} \geq \Delta$ for some specified positive constant Δ, and used a natural selection rule which selects the population yielding the largest sample mean value, based on a common sample of size n, as the best population. In this formulation, the sample size n should be determined so that the probability of a correct selection (PCS) will be at least P^* for a prespecified probability level $P^*(k^{-1} < P^* < 1)$, for all values of the specified Δ. Gupta (1956, 1965) imposed no restriction on the configurations of the parameters $\underset{\sim}{\theta}$, and proposed the subset selection approach to select a subset containing the best population with PCS $\geq P^*$, over all the possible configurations of the parameters $\underset{\sim}{\theta}$, where CS denotes the event that the best population is included in the selected subset. In this approach, the size of the selected subset is a random number, determined by the outcome of the experiment. In this approach no probability assessment is made regarding the quality of the other populations in the selected subsets. Later, Gupta and Santner (1973) and Santner (1975) introduced restricted subset selection procedures, in which the size of the selected subset is at most m, where $1 \leq m \leq k-1$. On the other hand, for the specified constants $\delta_2 > \delta_1 > 0$, Desu (1970) called a population π_i good if $\theta_{[k]} - \theta_i \leq \delta_1$, and bad if $\theta_{[k]} - \theta_i \geq \delta_2$. He studied the problem of selecting a subset of the populations so that none of the selected populations is bad. Lam (1986) studied this problem with a different dual selection goal, in which all the good populations should be included in the selected subset. For many other selection goals and formulation of the related selection

problems, the reader is referred to Gupta and Panchapakesan (1979).

In this paper, our aim is to control the quality of the selected populations. It is assumed that there is no prior information about the configurations of the parameter $\underset{\sim}{\theta}$. Hence, subset selection approach will be applied here. We desire that the best population be included in the selected subset, as well as, that each selected population be within some specified fixed distance from the best population.

This paper is organized as follows. We first formulate this selection problem in Section 2. Then we propose a restricted subset selection rule for general location-parameter models in Section 3. We also prove that this subset selection rule achieves the P^*-condition in the sense that the best is included in the selected subset as well as none of bad populations is selected. Special results regarding the normal distributions $N(\theta_i, \sigma^2)$ are presented in Section 4. In Section 5, modified subset selection rules are investigated. These modified selection rules will achieve the P^*-condition as well as control the expected value of the number of bad populations included in the selected subset. Finally, an example is used to illustrate the implementation of the selection rules.

2. Formulation of the Selection Problem

Let $\pi_1, \ldots, \pi_k$ denote $k(\geq 2)$ independent location-parameter models which have absolutely continuous cdf $G(x-\theta_i)$ with unknown location parameters $\theta_i, -\infty < \theta_i < \infty, i = 1, \ldots, k$. Let $\underset{\sim}{\theta} = (\theta_1, \ldots, \theta_k)$ and let $\theta_{[1]} \leq \ldots \leq \theta_{[k]}$ denote the ordered values of the parameters $\theta_1, \ldots, \theta_k$. It is assumed that the exact pairing between the ordered parameters and the unordered parameters is unknown. The population associated with the largest location parameter $\theta_{[k]}$ is called the best population. Let δ^* be a specified positive constant. Population π_i is said to be good if $\theta_{[k]} - \theta_i \leq \delta^*$ and bad otherwise. Let CS denote the event that the best population is included in the selected subset and only good populations are selected. Our goal is to derive a subset selection rule, say R, such that

$$P_{\underset{\sim}{\theta}}\{CS|R\} \geq P^* \quad \text{for all } \underset{\sim}{\theta} \in \Omega, \tag{2.1}$$

where $P^*(k^{-1} < P^* < 1)$ is a prespecified probability level and $\Omega = \{\underset{\sim}{\theta} = (\theta_1, \ldots, \theta_k) | -\infty < \theta_i < \infty, i = 1, \ldots, k\}$ is the whole parameter space.

Let $X_{ij}, j = 1, \ldots, n$ be a sample of size n from population π_i, and let $Y_i = Y(X_{i1}, \ldots, X_{in})$ be an appropriate statistic for the parameter θ_i. For example, in the normal distribution case, we let $Y_i = \overline{X}_i = \frac{1}{n} \sum_{j=1}^{n} X_{ij}$, the sample mean based on $(X_{i1}, \ldots, X_{in})$. It is assumed that Y_i also has an absolutely continuous cdf $F_n(y - \theta_i), i = 1, \ldots, k$. Also, let $Y_{[1]} \leq \ldots \leq Y_{[k]}$ denote the order statistics of $Y_1, \ldots, Y_k$ and let $\theta_{(i)}$ be the location parameter associated with the population $\pi_{(i)}$ which yields $Y_{[i]}$, and let $Y_{(i)}$ be the (unknown) statistic associated with the population $\pi_{[i]}$ having $\theta_{[i]}$ as its corresponding location parameter, $i = 1, \ldots, k$. Furthermore, we assume that the distribution function $F_n(\cdot)$ has the following property.

Assumption A: For each $c > 0$ and each fixed positive integer m, $\int_{-\infty}^{\infty} [F_n(y+c)]^m dF_n(y)$ is strictly increasing in n and tends to one as $n \to \infty$.

3. A Restricted Subset Selection Rule

For the given value δ^* and the prespecified probability level P^* let

$$n_0 = \min\{n \geq 1 | \int_{-\infty}^{\infty} [F_n(y + \delta^*/2)]^{k-1} dF_n(y) \geq P^*\}. \tag{3.1}$$

Under Assumption A, n_0 is well-defined and $n_0 < \infty$.

Let $Y_i = Y(X_{i1}, \ldots, X_{in_0}), i = 1, \ldots, k$. We propose a selection rule R_1 as follows:

$$R_1\text{: Select population } \pi_i \text{ if } Y_i \geq Y_{[k]} - \delta^*/2. \tag{3.2}$$

Let S denote the selected subset applying the selection rule R_1. That is, $S = \{\pi_i | Y_i \geq Y_{[k]} - \delta^*/2\}$. Then, according to the selection goal under study,

$$CS = \{\pi_{[k]} \in S \text{ and } \theta_{[k]} - \theta_i \leq \delta^* \text{ for all } \pi_i \in S\}. \tag{3.3}$$

Note that the form of the selection rule R_1 is similar to that of Gupta-type subset selection rule; see Gupta (1965). However, the problem involved here is to determine the sample size n to meet the P^*-condition; while in Gupta's subset selection rule, one is asked to determine the related critical value for the given P^* and sample size n.

Probability of a Correct Selection

Let $Z_{(i)} = Y_{(i)} - \theta_{[i]}, i = 1, \ldots, k$. Then $Z_{(1)}, \ldots, Z_{(k)}$ are iid, having the cdf $F_{n_0}(z)$. By the definition of n_0, analogous to Theorem 2.1 of Hsu (1981), one can obtain, for all $\underset{\sim}{\theta} \in \Omega$, that

$$\begin{aligned} P^* &\leq \int_{-\infty}^{\infty} [F_{n_0}(y + \delta^*/2)]^{k-1} dF_{n_0}(y) \\ &= P\{Z_{(k)} \geq Z_{(j)} - \delta^*/2, j = 1, \ldots, k-1\} \\ &= P_{\underset{\sim}{\theta}}\{Y_{(k)} \geq Y_{(j)} + \theta_{[k]} - \theta_{[j]} - \delta^*/2, j = 1, \ldots, k-1\} \\ &\leq P_{\underset{\sim}{\theta}}\{Y_{(k)} \geq Y_{(j)} - \delta^*/2, j = 1, \ldots, k-1, \text{ and } \theta_{[k]} - \theta_{[j]} \leq D_j, j = 1, \ldots, k\} \\ &\leq P_{\underset{\sim}{\theta}}\{\pi_{[k]} \in S \text{ and } \theta_{[k]} - \theta_j \leq D_j, \pi_j \in S\}. \end{aligned} \tag{3.4}$$

where $D_j = \max(\max_{i \neq j} Y_i - Y_j + \delta^*/2, 0)$.

Note that

$$\pi_j \in S \text{ iff } \max_{i \neq j} Y_i - Y_j + \delta^*/2 \leq \delta^*. \tag{3.5}$$

Combining (3.4) and (3.5) leads to the following: For all $\underset{\sim}{\theta} \in \Omega$,

$$\begin{aligned} P^* &\leq P_{\underset{\sim}{\theta}}\{\pi_{[k]} \in S \text{ and } \theta_{[k]} - \theta_j \leq D_j, \pi_j \in S\} \\ &\leq P_{\underset{\sim}{\theta}}\{\pi_{[k]} \in S \text{ and } \theta_{[k]} - \theta_j \leq \delta^*, \pi_j \in S\} \\ &= P_{\underset{\sim}{\theta}}\{CS|R_1\}. \end{aligned} \tag{3.6}$$

Least Favorable Configurations

For the given δ^*, let $\Omega_i(\delta^*) = \{\underset{\sim}{\theta} \in \Omega | \theta_{[k-i]} < \theta_{[k]} - \delta^* \leq \theta_{[k-i+1]}\}, i = 1, \ldots, k$, where $\theta_0 = -\infty$. That is, for $\underset{\sim}{\theta} \in \Omega_i(\delta^*)$, there are exact i good populations (including the best population). It is clear that $\Omega_1(\delta^*), \ldots, \Omega_k(\delta^*)$ are mutually exclusive and $\bigcup_{i=1}^{k} \Omega_i(\delta^*) = \Omega$.

For $\underset{\sim}{\theta} \in \Omega_k(\delta^*)$, all the k populations are good. Hence,

$$\begin{aligned} P_{\underset{\sim}{\theta}}\{CS|R_1\} &= P_{\underset{\sim}{\theta}}\{Y_{(k)} \geq \max_{1 \leq j \leq k} Y_j - \delta^*/2\} \\ &= P_{\underset{\sim}{\theta}}\{Y_{(k)} - \theta_{[k]} \geq Y_{(j)} - \theta[j] + \theta_{[j]} - \theta_{[k]} - \delta^*/2, j \neq k\} \\ &\geq P\{Z_{(k)} \geq Z_{(j)}) - \delta^*/2, j \neq k\} \\ &= \int_{-\infty}^{\infty} [F_{n_0}(z + \delta^*/2)]^{k-1} dF_{n_0}(z) \end{aligned} \tag{3.7}$$

for all $\underset{\sim}{\theta} \in \Omega_k(\delta^*)$. In (3.7), the inequality can be replaced by an equality when $\underset{\sim}{\theta} \in \Omega_k^\circ(\delta^*) = \{\underset{\sim}{\theta} \in \Omega_k(\delta^*) | \theta_1 = \ldots = \theta_k\}$. Therefore,

$$\begin{aligned} \inf_{\underset{\sim}{\theta} \in \Omega_k(\delta^*)} P_{\underset{\sim}{\theta}}\{CS|R_1\} &= \inf_{\underset{\sim}{\theta} \in \Omega_k^\circ(\delta^*)} P_{\underset{\sim}{\theta}}\{CS|R_1\} \\ &= \int_{-\infty}^{\infty} [F_{n_0}(z + \delta^*/2)]^{k-1} dF_{n_0}(z). \end{aligned} \tag{3.8}$$

For $\underset{\sim}{\theta} \in \Omega_1(\delta^*)$, all non-best populations are bad. Hence,

$$\begin{aligned} P_{\underset{\sim}{\theta}}\{CS|R_1\} &= P_{\underset{\sim}{\theta}}\{Y_{(i)} < Y_{(k)} - \delta^*/2, i \neq k\} \\ &= P_{\underset{\sim}{\theta}}\{Z_{(k)} > Z_{(i)} + \theta_{[i]} - \theta_{[k]} + \delta^*/2, i \neq k\} \\ &\geq P_{\underset{\sim}{\theta}}\{Z_{(k)} > Z_{(i)} - \delta^*/2, i \neq k\} \\ &= \int_{-\infty}^{\infty} [F_{n_0}(z + \delta^*/2)]^{k-1} dF_{n_0}(z), \end{aligned} \tag{3.9}$$

since $\theta_{[k]} - \theta_{[i]} > \delta^*$ for all $i \neq k$. In (3.9), the inequality can be replaced by an equality when $\underset{\sim}{\theta} \in \Omega_1^\circ(\delta^*) = \{\underset{\sim}{\theta} \in \Omega | \theta_{[1]} = \ldots = \theta_{[k-1]} = \theta_{[k]} - \delta^*\}$. Thus,

$$\begin{aligned} \inf_{\underset{\sim}{\theta} \in \Omega_1(\delta^*)} P_{\underset{\sim}{\theta}}\{CS|R_1\} &= \inf_{\underset{\sim}{\theta} \in \Omega_1^\circ(\delta^*)} P_{\underset{\sim}{\theta}}\{CS|R_1\} \\ &= \int_{-\infty}^{\infty} [F_{n_0}(z + \delta^*/2)]^{k-1} dF_{n_0}(z). \end{aligned} \tag{3.10}$$

For each $i = 2, \ldots, k-1$, on $\Omega_i(\delta^*)$, it is hard to find the corresponding least favorable configurations. However, we have the following result.

For each $\underset{\sim}{\theta} \in \Omega_i(\delta^*), i = 2, \ldots, k-1$,

$$\begin{aligned} & P_{\underset{\sim}{\theta}}\{CS|R_1\} \\ &= P_{\underset{\sim}{\theta}}\{Y_{(k)} \geq Y_{(j)} - \delta^*/2, j \neq k \text{ and } Y_{(\ell)} < Y_{[k]} - \delta^*/2, 1 \leq \ell \leq k-i\} \\ &\geq P_{\underset{\sim}{\theta}}\{Y_{(k)} \geq Y_{(j)} - \delta^*/2, j \neq k \text{ and } Y_{(\ell)} < Y_{(k)} - \delta^*/2, 1 \leq \ell \leq k-i\} \\ &= P_{\underset{\sim}{\theta}}\{Y_{(k)} \geq Y_{(j)} - \delta^*/2, k-i+1 \leq j \leq k-1, \text{ and } Y_{(k)} > Y_{(\ell)} + \delta^*/2, 1 \leq \ell \leq k-i\} \\ &= P_{\underset{\sim}{\theta}}\left\{ \begin{array}{c} Y_{(k)} - \theta_{[k]} \geq Y_{(j)} - \theta_{[j]} - \delta^*/2 + \theta_{[j]} - \theta_{[k]}, k-i+1 \leq j \leq k-1, \text{ and} \\ Y_{(k)} - \theta_{[k]} > Y_{(\ell)} - \theta_{[\ell]} + \delta^*/2 + \theta_{[\ell]} - \theta_{[k]}, 1 \leq \ell \leq k-i \end{array} \right\} \\ &\geq P_{\underset{\sim}{\theta}}\{Z_{(k)} \geq Z_{(j)} - \delta^*/2, k-i+1 \leq j \leq k-1, \text{ and } Z_{(k)} \geq Z_{(\ell)} - \delta^*/2, 1 \leq \ell \leq k-i\} \\ &= \int_{-\infty}^{\infty} [F_{n_0}(z + \delta^*/2)]^{k-1} dF_{n_0}(z). \end{aligned} \tag{3.11}$$

In (3.11), the second inequality is obtained from the fact that for $\underset{\sim}{\theta} \in \Omega_i(\delta^*), \theta_{[j]} - \theta_{[k]} \leq 0$ for $k-i+1 \leq j \leq k-1$ and $\theta_{[\ell]} - \theta_{[k]} < -\delta^*$ for $1 \leq \ell \leq k-i$. The second inequality can be replaced by an equality when $\underset{\sim}{\theta} = (\theta_1, \ldots, \theta_k)$ is such that $\theta_{[j]} = \theta_{[k]}, k-i+1 \leq j \leq k$, and $\theta_{[\ell]} = \theta_{[k]} - \delta^*, 1 \leq \ell \leq k-i$.

Based on the preceding discussions, we conclude that

$$\begin{aligned} \inf_{\underset{\sim}{\theta} \in \Omega} P_{\underset{\sim}{\theta}}\{CS|R_1\} &= \inf_{\underset{\sim}{\theta} \in \Omega'} P_{\underset{\sim}{\theta}}\{CS|R_1\} \\ &= \int_{-\infty}^{\infty} [F_{n_0}(z+\delta^*/2)]^{k-1} dF_{n_0}(z), \end{aligned} \tag{3.12}$$

where $\Omega' = \Omega_1^\circ(\delta^*) \cup \Omega_k^\circ(\delta^*)$.

4. Selection of the Best Normal Population

Let $X_{ij}, j = 1, \ldots, n$, be a sample of size n from $N(\theta_i, \sigma^2), i = 1, \ldots, k$, where the common variance σ^2 may be either known or unknown. The best population is the one associated with the largest $\theta_{[k]}$. We consider two situations according to whether the common variance σ^2 is known or unknown.

4.1 Selection Rule for σ^2 Known Case

Let $Z_1, \ldots, Z_k$ denote k iid random variables, having $N(0,1)$ distribution. For the given P^*, let d^* be the solution such that

$$P\{Z_k \geq Z_i - d^*, i \neq k\} = P^*. \tag{4.1}$$

Note that the value of d^* should be positive since $P^* > k^{-1}$. Also, for certain given k and P^* values, the corresponding d^* values have been tabulated by Gupta, Nagel and Panchapakesan (1973), and Gupta, Panchapakesan and Sohn (1985), among many others.

For given values of δ^*, σ and P^*, we determine the sample size n_0 by

$$n_0 = < 4d^{*2}\sigma^2/\delta^{*2} > \tag{4.2}$$

where $< x >$ denotes the smallest integer not less than x. Let $Y_i = \overline{X}_i = \frac{1}{n_0} \sum_{j=1}^{n_0} X_{ij}$ be the sample mean based in a sample of size n_0 taken from population $\pi_i, i = 1, \ldots, k$. We propose the selection rule R_2 given as follows:

$$R_2\text{: Select population } \pi_i \text{ if } \overline{X}_i \geq \max_{1 \leq j \leq k} \overline{X}_j - \frac{d^*\sigma}{\sqrt{n_0}}. \tag{4.3}$$

Let $S = \{\pi_i | \overline{X}_i \geq \max_{1\leq j\leq k} \overline{X}_j - d^*\sigma/\sqrt{n_0}\}$, the selected subset employing the selection rule R_2. Then, according to the selection goal,

$$CS = \{\pi_{[k]} \in S \text{ and } \theta_{[k]} - \theta_i \leq \delta^* \text{ for all } \pi_i \in S\}.$$

Following (2.9) of Hsu (1981), for all $\underset{\sim}{\theta} \in \Omega$, one can obtain

$$\begin{aligned} P^* &\leq P_{\underset{\sim}{\theta}}\{\overline{X}_{(k)} \geq \max_{1\leq j\leq k} \overline{X}_j - \frac{d^*\sigma}{\sqrt{n_0}} \text{ and } \theta_{[k]} - \theta_i \leq D_i, i = 1, \ldots, k\} \\ &\leq P_{\underset{\sim}{\theta}}\{\pi_{[k]} \in S \text{ and } \theta_{[k]} - \theta_i \leq D_i, \pi_i \in S\}, \end{aligned} \tag{4.4}$$

where $D_i = \max(\max_{j\neq i} \overline{X}_j - \overline{X}_i + \frac{d^*\sigma}{\sqrt{n_0}}, 0)$.

Note that $\pi_i \in S$ iff $D_i \leq 2d^*\sigma/\sqrt{n_0}$. Also, $2d^*\sigma/\sqrt{n_0} \leq \delta^*$, which is obtained from (4.2). Therefore, for all $\underset{\sim}{\theta} \in \Omega$,

$$\begin{aligned} P^* &\leq P_{\underset{\sim}{\theta}}\{\pi_{[k]} \in S \text{ and } \theta_{[k]} - \theta_i \leq D_i, \pi_i \in S\} \\ &\leq P_{\underset{\sim}{\theta}}\{\pi_{[k]} \in S \text{ and } \theta_{[k]} - \theta_i \leq \delta^*, \pi_i \in S\} \\ &= P_{\underset{\sim}{\theta}}\{CS|R_2\}. \end{aligned} \tag{4.5}$$

4.2 Selection Rule for the Unknown σ^2 Case

When the value of the common variance σ^2 is unknown, we propose a two-stage selection rule as follows.

First, take $n_0(\geq 2)$ observations from each of the k populations. Compute $\overline{X}_i(1) = \frac{1}{n_0}\sum_{j=1}^{n_0} X_{ij}$ and $W = \left[\frac{1}{k(n_0-1)}\sum_{i=1}^{k}\sum_{j=1}^{n_0}(X_{ij} - \overline{X}_i(1))^2\right]^{1/2}$. Note that $\overline{X}_i(1), \ldots, \overline{X}_k(1)$ and W are mutually independent and $k(n_0-1)W^2/\sigma^2$ follows a chi-square distribution having degrees of freedom $k(n_0 - 1)$. For the given P^*, let c^* be the solution of the following equation:

$$P\{Z_k \geq Z_i - c^*V, i \neq k\} = P^*, \tag{4.6}$$

where $k(n_0 - 1)V^2$ has a chi-square distribution with $k(n_0 - 1)$ degrees of freedom, and is independent of $(Z_1, \ldots, Z_k)$. The value of c^* should be positive since $P^* > k^{-1}$ and $V \geq 0$. For certain given P^*, k and n_0 values, the corresponding c^* values can be found in Gupta, Panchapakesan and Sohn (1985). Let

$$N_0 = \max\{n_0, < \frac{4c^{*2}W^2}{\delta^{*2}} >\}. \tag{4.7}$$

Secondly, take $N_0 - n_0$ additional observations from each of the k populations if necessary. Compute $\overline{X}_i = \frac{1}{N_0}\sum_{j=1}^{N_0} X_{ij}, i = 1, \ldots, k$. Then, the two-stage subset selection rule R_3 is defined by

$$R_3\text{: Select population } \pi_i \text{ if } \overline{X}_i \geq \max_{1\leq j\leq k} \overline{X}_j - \frac{c^*W}{\sqrt{N_0}}. \tag{4.8}$$

Let S denote the selected subset applying the selection rule R_3. Then, $S = \{\pi_i | \overline{X}_i \geq \max_{1\leq j\leq k} \overline{X}_j - c^*W/\sqrt{N_0}\}$, and

$$CS = \{\pi_{[k]} \in S \text{ and } \theta_{[k]} - \theta_i \leq \delta^*, \pi_i \in S\}.$$

Following (2.12) of Hus (1981), for all $\underset{\sim}{\theta} \in \Omega$, we can obtain

$$\begin{aligned} P^* &\leq P_{\underset{\sim}{\theta}}\{\overline{X}_{(k)} \geq \max_{1\leq j\leq k} \overline{X}_j - \frac{c^*W}{\sqrt{N_0}}, \text{ and } \theta_{[k]} - \theta_i \leq D_i^*, i = 1, \ldots, k\} \\ &\leq P_{\underset{\sim}{\theta}}\{\pi_{[k]} \in S \text{ and } \theta_{[k]} - \theta_i \leq D_i^*, \pi_i \in S\}, \end{aligned} \tag{4.9}$$

$$\text{where } D_i^* = \max\left(\max_{j\neq i} \overline{X}_j - \overline{X}_i + \frac{c^*W}{\sqrt{N_0}}, 0\right). \tag{4.10}$$

Since $\pi_i \in S$ iff $D_i^* \leq 2c^*W/\sqrt{N_0} \leq \delta^*$, where the second inequality is obtained from the definition of N_0, see (4.7), we can conclude that for all $\underset{\sim}{\theta} \in \Omega$,

$$\begin{aligned} P^* &\leq P_{\underset{\sim}{\theta}}\{\pi_{[k]} \in S \text{ and } \theta_{[k]} - \theta_i \leq D_i^*, \pi_i \in S\} \\ &\leq P_{\underset{\sim}{\theta}}\{\pi_{[k]} \in S \text{ and } \theta_{[k]} - \theta_i \leq \delta^*, \pi_i \in S\} \\ &= P_{\underset{\sim}{\theta}}\{CS | R_3\}. \end{aligned} \tag{4.11}$$

5. A Modified Selection Goal

Even though one can choose high probability P^* to guarantee that all the selected populations are good, there is still some possibility that some bad populations may be included in the selected subset. Let T denote the number of bad populations included in the selected subset. Then $0 \leq T \leq k-1$. We desire that the expected value of T be small. That is, additional to the P^*-condition, we would like to impose the following requirement on the selection rules:

$$\max_{\underset{\sim}{\theta}\in\Omega} E_{\underset{\sim}{\theta}}[T] \leq q, \tag{5.1}$$

where q is a specified value such that $0 < q < k-1$. Usually, the value of q is chosen to be small. Note that

$$\max_{\underset{\sim}{\theta}\in\Omega} E_{\underset{\sim}{\theta}}[T] = \max_{1\le i\le k-1} \max_{\underset{\sim}{\theta}\in\Omega_i(\delta^*)} E_{\underset{\sim}{\theta}}[T]. \tag{5.2}$$

Before we go further to derive selection rules satisfying both the P^*-condition and the requirement of (5.1), we first give a representation for $\max_{\underset{\sim}{\theta}\in\Omega} E_{\underset{\sim}{\theta}}[T]$. In the following, all discussions and results pertain to the normal distribution models.

5.1 The Case When σ^2 is Known

When σ^2 is known, we let $R_2(n)$ be a subset selection rule which has the same form as that of R_2, and is based on a sample of size n from each of the k populations. That is,

$$R_2(n)\text{: Select population } \pi_i \text{ if } \overline{X}_i \ge \max_{1\le j\le k} \overline{X}_j - \frac{d^*\sigma}{\sqrt{n}}, \tag{5.3}$$

where $\overline{X}_i = \frac{1}{n}\sum_{j=1}^{n} X_{ij}$ and d^* is the solution of (4.1).

let $E_{\underset{\sim}{\theta}}[T|R_2(n)]$ denote the expected value of T when selection rule $R_2(n)$ is applied. Then, for each $\underset{\sim}{\theta} \in \Omega_i(\delta^*), 1 \le i \le k-1$,

$$E_{\underset{\sim}{\theta}}[T|R_2(n)] = \sum_{j=1}^{k-i} \int_{-\infty}^{\infty} \prod_{\substack{m=1\\ m\ne j}}^{K} \Phi\left(z + d^* - \frac{\sqrt{n}(\theta_{[m]} - \theta_{[j]})}{\sigma}\right) \phi(z)dz, \tag{5.4}$$

where $\Phi(\cdot)$ and $\phi(\cdot)$ denote, respectively, the cdf and pdf of a standard normal distribution. We ahve the following lemma.

Lemma 5.1 $\sup_{\underset{\sim}{\theta}\in\Omega_i(\delta^*)} E_{\underset{\sim}{\theta}}[T|R_2(n)] = (k-i)H(d^*, \delta^*, n, \sigma)$, where

$$H(d^*, \delta^*, n, \sigma) = \int_{-\infty}^{\infty} [\Phi(z+d)]^{k-2}\Phi(z + d - \frac{\sqrt{n}\delta^*}{\sigma})d\Phi(z). \tag{5.5}$$

Proof: First for any $\underset{\sim}{\theta} \in \Omega_i(\delta^*)$, keep $\theta_{[1]}, \ldots, \theta_{[k-i]}$ and $\theta_{[k]}$ being fixed. From (5.4), we see that $E_{\underset{\sim}{\theta}}[T|R_2(n)]$ is decreasing in $\theta_{[k-i+1]}, \ldots, \theta_{[k-1]}$.

Next, let $\theta_{[k-i+1]}, \ldots, \theta_{[k]}$ be fixed. Suppose that the $h(1 \le h \le k-i)$ smallest parameters $\theta_{[j]}(1 \le j \le h)$ are equal and the common value is denoted by θ. We want to show that $E_{\underset{\sim}{\theta}}[T|R_2(n)]$ is increasing in θ for $\theta \le \min(\theta_{[h+1]}, \theta_{[k]} - \delta^*)$.

When $\theta_{[1]} = \ldots = \theta_{[h]} = \theta$, from (5.4),

$$E_{\underset{\sim}{\theta}}[T|R_2(n)] = h\int_{-\infty}^{\infty}[\Phi(z+d)]^{h-1}\prod_{m=h+1}^{K}\Phi(z+d-\frac{\sqrt{n}(\theta_{[m]}-\theta)}{\sigma})d\Phi(z)$$
$$+\sum_{j=h+1}^{k-i}\int_{-\infty}^{\infty}\left[\Phi\left(z+d-\frac{\sqrt{n}(\theta-\theta_{[j]})}{\sigma}\right)\right]^{h}\prod_{\substack{m=h+1\\m\neq j}}^{k}\Phi\left(z+d-\frac{\sqrt{n}(\theta_{[m]}-\theta_{[j]})}{\sigma}\right)d\Phi(z),$$

where $\sum_{x=a}^{b} \equiv 0$ if $a > b$. Then for $\theta \leq \min(\theta_{[h+1]}, \theta_{[k]} - \delta^*)$,

$$\frac{dE_{\underset{\sim}{\theta}}[T|R_2(n)]}{d\theta}$$
$$= \frac{\sqrt{n}h}{\sigma}\sum_{j=h+1}^{k}\int_{-\infty}^{\infty}[\Phi(z+d)]^{h-1}\left[\prod_{\substack{m=h+1\\m\neq j}}^{k}\Phi\left(z+d-\frac{\sqrt{n}(\theta_{[m]}-\theta)}{\sigma}\right)\right]$$
$$\phi\left(z+d-\frac{\sqrt{n}(\theta_{[j]}-\theta)}{\sigma}\right)d\Phi(z)$$
$$-\frac{\sqrt{n}h}{\sigma}\sum_{j=h+1}^{k-i}\int_{-\infty}^{\infty}\left[\Phi(z+d-\frac{\sqrt{n}(\theta-\theta_{[j]})}{\sigma})\right]^{h-1}\phi\left(z+d-\frac{\sqrt{n}(\theta-\theta_{[j]})}{\sigma}\right)$$
$$\left[\prod_{\substack{m=h+1\\m\neq j}}^{k}\Phi\left(z+d-\frac{\sqrt{n}(\theta_{[m]}-\theta_{[j]})}{\sigma}\right)\right]d\Phi(z)$$
$$= \frac{\sqrt{n}h}{\sigma}\sum_{j=k-i+1}^{k}\int_{-\infty}^{\infty}[\Phi(z+d)]^{h-1}\left[\sum_{\substack{m=h+1\\m\neq j}}^{k}\Phi\left(z+d-\frac{\sqrt{n}(\theta_{[m]}-\theta)}{\sigma}\right)\right]$$
$$\phi\left(z+d-\frac{\sqrt{n}(\theta_{[j]}-\theta)}{\sigma}\right)d\Phi(z)$$
$$+\frac{\sqrt{n}h}{\sigma}\sum_{j=h+1}^{k-i}\int_{-\infty}^{\infty}[\Phi(z+d)]^{h-1}\left[\sum_{\substack{m=h+1\\m\neq j}}^{k}\Phi\left(z+d-\frac{\sqrt{n}(\theta_{[m]}-\theta)}{\sigma}\right)\right]$$
$$\times\left[\phi\left(z+d-\frac{\sqrt{n}(\theta_{[j]}-\theta)}{\sigma}\right)\phi(z)-\phi(z+d)\phi\left(z+\frac{\sqrt{n}(\theta-\theta_{[j]})}{\sigma}\right)\right]dz$$
$$\geq 0,$$

since $\phi\left(z+d-\frac{\sqrt{n}(\theta_{[j]}-\theta)}{\sigma}\right)\phi(z)-\phi(z+d)\phi\left(z+\frac{\sqrt{n}(\theta-\theta_{[j]})}{\sigma}\right) \geq 0$ for $\theta \leq \min(\theta_{[h+1]}, \theta_{[k]} - \delta^*)$ for all $j \geq h+1$. Hence,

$$\sup_{\underset{\sim}{\theta}\in\Omega_i(\delta^*)} E_{\underset{\sim}{\theta}}[T|R_2(n)] = \lim_{\substack{\theta^*_{[j]}\to\theta^*_{[k]}-\delta^*\\1\leq j\leq k-i}} E_{\underset{\sim}{\theta}^*}[T|R_2(n)]$$
$$= (k-i)H(d^*,\delta^*,n,\sigma),$$

where $\underset{\sim}{\theta}^* = (\theta_1^*, \ldots, \theta_k^*)$ is such that $\theta^*_{[k-i]} < \theta^*_{[k-i+1]} = \ldots = \theta^*_{[k-1]} = \theta^*_{[k]} - \delta^*$. □

By Lemma 5.1 and (5.2),

$$\max_{\underset{\sim}{\theta}\in\Omega} E_{\underset{\sim}{\theta}}[T|R_2(n)] = (k-1)H(d^*, \delta^*, n, \sigma). \tag{5.6}$$

For a given $q, (0 < q < k-1)$, let

$$n^* = \min\{n \geq n_0 | (k-1)H(d^*, \delta^*, n, \sigma) \leq q\}. \tag{5.7}$$

Since $H(d^*, \delta^*, n, \sigma)$ is decreasing in n and tends to zero as $n \to \infty$, n^* is well-defined and $n^* < \infty$.

Based on the preceding discussions, one can see that the subset selection rule $R_2(n^*)$ satisfies both the P^*-condition and the requirement of (5.1).

5.2 The Case When σ^2 Is Unknown

For any given P^*, let c^* be the solution of (4.6). For any given q, let

$$b^* = \inf\{b | G(b) \leq q/(k-1)\} \tag{5.8}$$

where

$$G(b) = P\{Z_1 > Z_j - c^*V, 2 \leq j \leq k-1, \text{ and } Z_1 > Z_k + (b-c^*)V\}, \tag{5.9}$$

where $Z_1, \ldots, Z_k$ and V are those defined in Section 4. Note that $G(0) = P^*, G(b)$ is decreasing in b and $\lim_{b\to\infty} G(b) = 0$. Hence, for q being small enough, b^* is well-defined and $c^* < b^* < \infty$.

To satisfy the additional requirement of (5.1), we modify the selection rule R_3 by redetermining the sample size N_0 by

$$N^* = \max\left\{n_0, < \frac{W^2}{\delta^{*2}} \max(4c^{*2}, b^{*2}) >\right\}. \tag{5.10}$$

That is, at the second stage, we take $N^* - n_0$ additional observations from each of the k populations if necessary. Compute $\overline{X}_i = \frac{1}{N^*} \sum_{j=1}^{N^*} X_{ij}, i = 1, \ldots, k$. Then, include population π_i in the selected subset if $\overline{X}_i \geq \max_{1\leq j\leq k} \overline{X}_j - \frac{c^*W}{\sqrt{N^*}}$. We note this modifed two-stage subset selection rule by R_3^*. Since $N^* \geq N_0$, the P^*-condition is always satisfied. It suffices to show that $\max_{\underset{\sim}{\theta}\in\Omega} E[T|R_3^*] \leq q$ for small q.

Let $C_n = \{N^* = n\}, n \geq n_0$. For each $\underset{\sim}{\theta} \in \Omega_i(\delta^*), 1 \leq i \leq k-1$,

$$
\begin{aligned}
& E_{\underset{\sim}{\theta}}[T|R_3^*] \\
= & \sum_{j=1}^{k-i} P_{\underset{\sim}{\theta}} \left\{ \overline{X}_j > \overline{X}_m - \frac{c^* W}{\sqrt{N^*}}, m \neq j \right\} \\
= & \sum_{j=1}^{k-i} P \left\{ Z_j > Z_m + \frac{\sqrt{N^*}(\theta_{[m]} - \theta_{[j]})}{\sigma} - c^* V, m \neq j \right\} \qquad (5.11) \\
= & \sum_{j=1}^{k-i} \sum_{n=n_0}^{\infty} \int_{C_n} \int_{-\infty}^{\infty} \prod_{\substack{m=1 \\ m \neq j}}^{k} \Phi \left(z + c\nu - \frac{\sqrt{n}(\theta_{[m]} - \theta_{[j]})}{\sigma} \right) d\Phi(z) dF_V(\nu) \\
= & \sum_{n=n_0}^{\infty} \int_{C_n} E_{\underset{\sim}{\theta}}[T|(v,n)] dF_V(\nu),
\end{aligned}
$$

where $E_{\underset{\sim}{\theta}}[T|(v,n)] = \sum_{j=1}^{k-1} \int_{-\infty}^{\infty} \prod_{\substack{m=1 \\ m \neq j}}^{k} \Phi \left(z + c\nu - \frac{\sqrt{n}(\theta_{[m]} - \theta_{[j]})}{\sigma} \right) d\Phi(z)$, $V = W/\sigma$, which is independent of $(Z_1, \ldots, Z_k)$, $k(n_0 - 1)V^2$ has a chi-square distribution with $k(n_0 - 1)$ degrees of freedom, and $F_V(\cdot)$ denotes the distribution function of V.

Analogous to the proof of Lemma 5.1, one can obtain that

$$
E_{\underset{\sim}{\theta}}[T|(v,n)] \leq (k-i) \int_{-\infty}^{\infty} [\Phi(z + c^* v)]^{k-2} \Phi \left(z + c^* v - \frac{\sqrt{n}\delta^*}{\sigma} \right) d\Phi(z).
$$

Therefore,

$$
\begin{aligned}
& E_{\underset{\sim}{\theta}}[T|R_3^*] \\
\leq & \sum_{n=n_0}^{\infty} \int_{C_n} (k-i) \int_{-\infty}^{\infty} [\Phi(z + c^* v)]^{k-2} \Phi \left(z + c^* v - \frac{\sqrt{n}\delta^*}{\sigma} \right) d\Phi(z) dF_V(v) \\
= & (k-i) P \left\{ Z_1 > Z_j - c^* V, 2 \leq j \leq k-1, \text{ and } Z_1 > Z_k - c^* V + \frac{\sqrt{N^*}\delta^*}{\sigma} \right\} \\
\leq & (k-i) P \left\{ Z_1 > Z_j - c^* V, 2 \leq j \leq k-1, \text{ and } Z_1 > Z_k + (b^* - c^*) V \right\},
\end{aligned}
$$

where the last inequality is obtained due to the definition of N^*, see (5.10).

Hence,

$$
\begin{aligned}
\max_{\underset{\sim}{\theta} \in \Omega} E_{\underset{\sim}{\theta}}[T|R_3^*] &= \max_{1 \leq i \leq k-1} \max_{\underset{\sim}{\theta} \in \Omega_i(\delta^*)} E_{\underset{\sim}{\theta}}[T|R_3^*] \\
&\leq (k-1) P\{Z_1 > Z_j - c^* V, 2 \leq j \leq k-1, Z_1 > Z_k + (b^* - c^*) V\} \\
&\leq q
\end{aligned}
$$

which is guaranteed by (5.8).

6. **An Illustrative Example** (The data is taken from Example 3, page 506, of Gupta and Panchapakesan (1979)).

An experimenter wants to compare the glowing time of five different types of phosphorescent coatings of airplane instrument dials. Assume that the distributions of the glowing time for each type of phosphorescent coatings are normal with a common unknown variance σ^2. The experimenter is interested in the selection of the best among the five types of phosphorescent coatings of airplane instruments dials. However, he has no idea about the possible configurations of the parameters. Hence subset selection approach is displayed here. He desires that not just the best will be selected, but also each selected population should be within the $\delta^* = 10$ distance from the unknown best.

He chooses $P^* = 0.90$ and $n_0 = 5$. The coated dials were then excited with an ultraviolet light. The upper part of Table 1 shows the number of minutes each dial glows after the light source was turned off.

Table 1. Glowing Time of Five Types of Phosphorescent Coatings

	Coatings				
	1	2	3	4	5
	45.7	51.7	45.9	54.8	65.9
observations	48.4	46.4	54.8	55.6	65.4
taken at the	51.9	49.8	62.9	63.5	60.0
first-stage	57.0	52.7	64.7	61.6	70.1
	41.0	48.1	54.3	55.7	69.5
n_0	5	5	5	5	5
$\overline{X}_i(1)$	48.8	49.74	56.52	58.24	66.18
	$W^2 = \frac{1}{k(n_0-1)} \sum_{i=1}^{k} \sum_{j=1}^{n_0} (X_{ij} - \overline{X}_i(1))^2 = 26.7305$				
observations	61.4	54.8	57.9	59.2	64.0
taken at the	47.0	54.0	53.9	53.2	56.0
second-stage	56.9	44.7	47.9	66.7	61.4
N_0	8	8	8	8	8
$\overline{X}_i$	51.1625	50.275	55.2875	58.7875	64.0375

For $k = 5, n_0 = 5, P^* = 0.90$, from Gupta, Panchapakesan and Sohn (1985), $c^* = 1.92727\sqrt{2}$. Therefore $N_0 = \max\left\{n_0, < \frac{4c^{*2}W^2}{\delta^{*2}} >\right\} = 8$. Hence $N_0 - n_0 = 3$ additional observations should be taken from each population. The observations taken at the second-stage are given in the lower part of Table 1.

According to the selection rule R_3, we include populations π_i in the selected subset if $\overline{X}_i \geq \max_{1 \leq j \leq 5} \overline{X}_j - \frac{c^* W}{\sqrt{N_0}}$, where $\frac{c^* W}{\sqrt{N_0}} = \frac{1.92727\sqrt{2} \times \sqrt{26.7305}}{\sqrt{8}} = 5.0149$. Hence only population π_5 is selected. From (4.10), $D_5^* = 0$. Hence, we can state with at least 90% confidence that population π_5 is the best population, since $\theta_{[k]} - \theta_5 \leq D_5^* = 0$.

References

Bechhofer, R.E. (1954). A single-sample multiple decision procedure for ranking means of normal populations with known variances. *Ann. Math. Statist.*, **25**, 16–39.

Desu, M.N. (1970). A selection problem. *Ann. Math. Statist.*, **41**, 1596–1603.

Gupta, S.S. (1956). On a decision rule for a problem in ranking means. Inst. Statist. Mimeo. Series 150, University of North Carolina, Chapel Hill, NC.

Gupta, S.S. (1965). On some multiple decision (selection and ranking) rules. *Technometrics*, **7**, 225–245.

Gupta, S.S., Nagel, K. and Panchapakesan, S. (1973). On the order statistics from equally correlated normal random variables. *Biometrika*, **60**, 403–413.

Gupta, S.S. and Panchapakesan, S. (1979). *Multiple Decision Procedures*, New York: John Wiley.

Gupta, S.S., Panchapakesan, S., and Sohn, J.K. (1985). On the distribution of the studentized maximum of equally correlated normal random variables. *Commun. Statist. - Simula. Computa.*, **14 (1)**, 103–135.

Gupta, S.S. and Santner, T.J. (1973). On selection and ranking procedures – a restricted subset selection rule. *Proceedings of the 39th Session of the International Statistical Institute*, Vienna, Austria, **1**, 409–417.

Hsu, J.C. (1981). Simultaneous confidence intervals for all distances from the "best". *Ann. Statist.*, **9**, 1026–1034.

Lam, K. (1986). A new procedure for selecting good populations. *Biometrika*, **73**, 201–206.

Santner, T.J. (1975). A restricted subset selection approach to ranking and decision problem. *Ann. Statist*, **3**, 334–349.